ani Slučajni Uzorak: Teorijski Pristup i Primjena

Kristina Zogović

Stratifikovani Slučajni Uzorak u Analizi Podataka

Podnaslov: Teorijski pristup, pregled literature, razlozi za upotrebu, prednosti, mane i case study

Autor: Kristina Zogović

Dopunski rad
Datum: 2025-09-01

Sadržaj

Teorijski pristup stratifikovanom uzorku

- Definicija: Stratifikovani slučajni uzorak je metoda uzorkovanja gde se populacija dijeli u međusobno isključive i kolektivno iscrpne grupe (strate) na osnovu ključnih karakteristika, a zatim se iz svake strate slučajno bira uzorak.

Teorijski pristup stratifikovanom uzorku

- Osnovni koraci:
  - Formiranje strata: Dijeljenje populacije na homogene podgrupe (npr. po vremenskim periodima, zemljama ili kategorijama).
  - Alokacija uzorka: Proporcionalna (prema veličini strate) ili optimalna (Neymanova, uzimajući varijansu).
  - Slučajno uzorkovanje unutar strata.
  - Procjena parametara: Ponderisani prosjeci za populacioni prosjek.

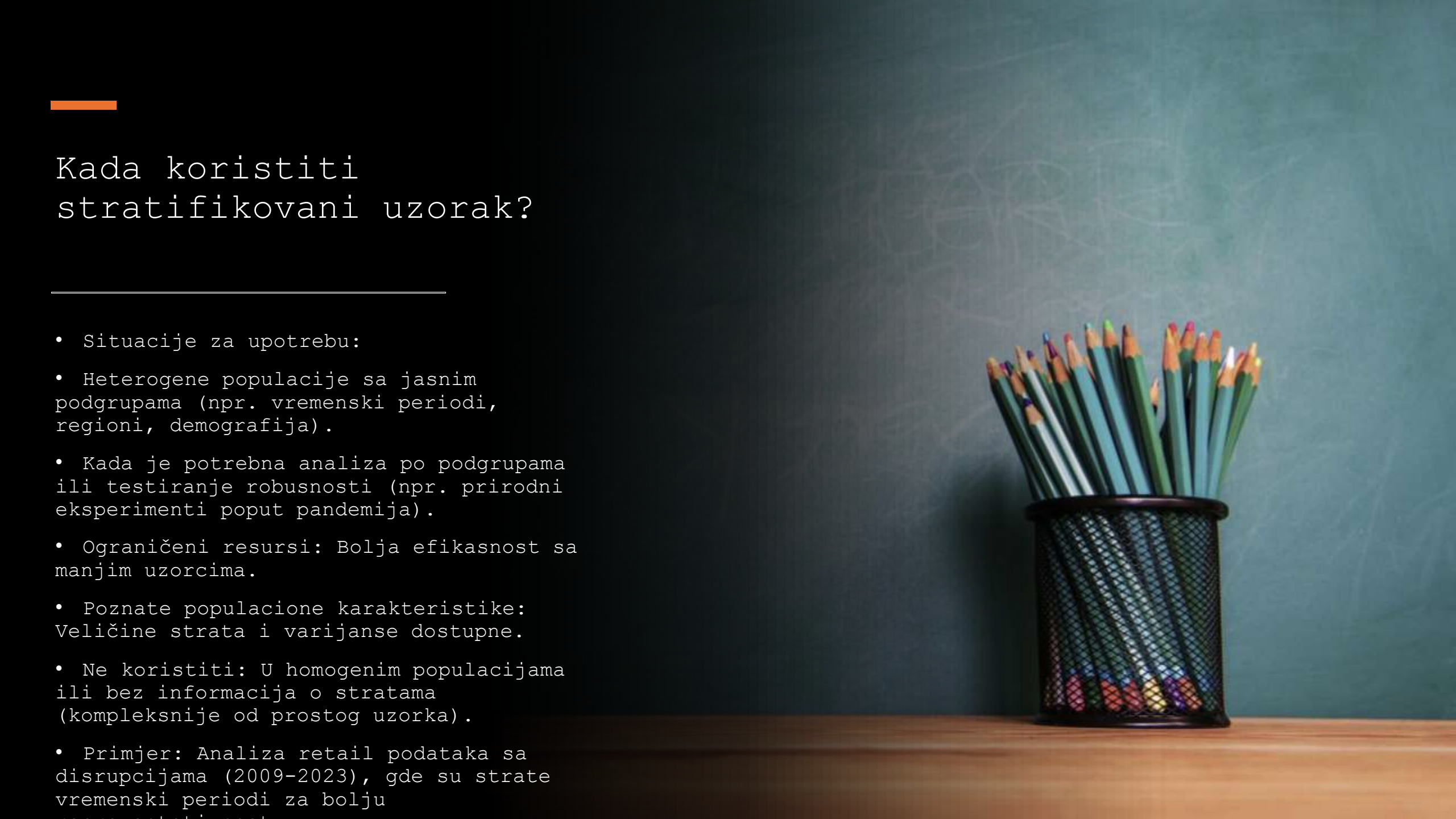
- Cilj: Smanjenje varijanse procene i poboljšanje reprezentativnosti u heterogenim populacijama.

Zašto koristiti stratifikovani uzorak?

- Glavni razlozi:
  - Povećava preciznost procjena u heterogenim populacijama smanjenjem varijanse.
  - Osigurava reprezentativnost ključnih podgrupa (strata), sprečavajući pristrasnost.
  - Omogućava zasebnu analizu podgrupa (subgroup analysis), korisno za poređenja.
  - Efikasniji od prostog slučajnog uzorka kada postoje poznate razlike u populaciji.
  - Primer: U retail analizi, zašto? Zbog heterogenosti po vremenskim periodima (npr. krize vs. rast), što omogućava bolje razumijevanje disrupcija poput COVID-19.

Kada koristiti stratifikovani uzorak?

- Situacije za upotrebu:
- Heterogene populacije sa jasnim podgrupama (npr. vremenski periodi, regioni, demografija).
- Kada je potrebna analiza po podgrupama ili testiranje robusnosti (npr. prirodni eksperimenti poput pandemija).
- Ograničeni resursi: Bolja efikasnost sa manjim uzorcima.
- Poznate populacione karakteristike: Veličine strata i varijanse dostupne.
- Ne koristiti: U homogenim populacijama ili bez informacija o stratama (kompleksnije od prostog uzorka).
- Primjer: Analiza retail podataka sa disrupcijama (2009-2023), gde su strate vremenski periodi za bolju



Prednosti stratifikovanog uzorka

- Prednosti:
  - Veća preciznost: Smanjena varijansa procjene (do 35-40% bolje od prostog uzorka po simulacijama).
  - Bolja reprezentativnost: Svaka strata je proporcionalno uključena, smanjujući bias.
  - Omogućava podgrupske analize: Nezavisne procjene po stratama.
  - Efikasnost: Optimalna alokacija minimizuje troškove uzorkovanja.
  - Robusnost: Bolje rukovanje heterogenošću, npr. u ekonomskim analizama sa krizним periodima.

Mane stratifikovanog uzorka

- Kompleksnost: Zahtijeva prethodno znanje o populaciji za formiranje strata.
- Potreba za informacijama: Veličine strata i varijanse moraju biti poznate, što može biti izazov.
- Veći troškovi pripreme: Dizajn i implementacija su složeniji od prostog uzorka.
- Rizik grešaka: Nepravilno formiranje strata može dovesti do biasa ili veće varijanse.
- Ograničenja: Manje efikasno u malim populacijama ili bez jasnih podgrupa.

Case Study: Primjena u analizi retail podataka (2009-2023)

- Uvod u case study: Primjena stratifikovanog uzorka na analizu vremenskih disrupcija i tržišne otpornosti u retail sektoru.
- Kontekst: Heterogena populacija sa periodima: Post-Crisis (2009-2012), Stable Growth (2013-2019), COVID-19 (2020-2021), Recovery (2022-2023).
- Zašto stratifikacija? Heterogenost po periodima, COVID-19 kao test robusnosti, analiza po podgrupama (zemlje, proizvodi).

Kreiranje i priprema podataka u case study

Karakteristike Dataset-a

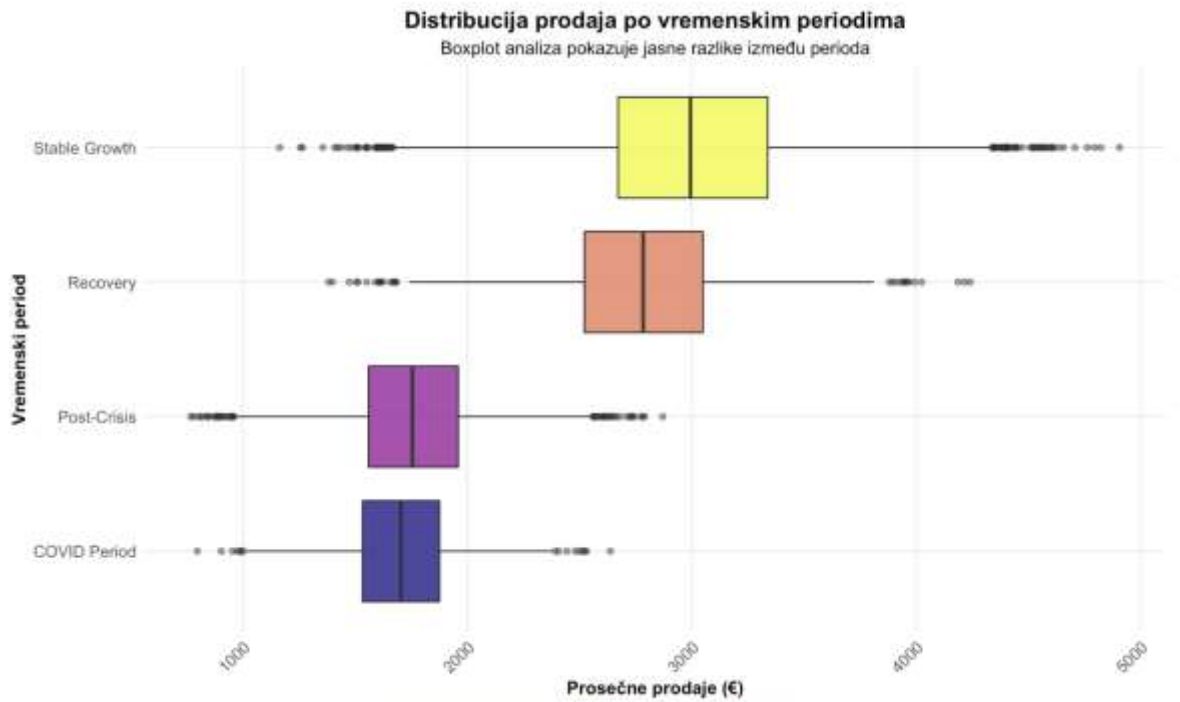
Ukupne transakcije: 25,005 **Broj zemalja:** 6 **Kategorije proizvoda:** 5 **Vremenski period:** 15 godina

Deskriptivne statistike po vremenskim periodima

period	Broj transakcija	Prosečne prodaje (€)	Medijan prodaja (€)	Std. devijacija	Min (€)	Max (€)
COVID Period	3300	1707	1704	251	797	2638
Post-Crisis	6580	1757	1755	299	769	2871
Recovery	3316	2790	2784	396	1382	4240
Stable Growth	11809	3002	2995	502	1164	4906

Eksplorativna analiza u case study

Ključni uvidi iz eksplorativne analize:
- **Stable Growth** period ima najviše prosečne prodaje (~€3,002)
- **COVID Period** pokazuje drastičan pad (~€1,706)
- **Recovery** period pokazuje znakove oporavka (~€2,787)
- Jasno vidljiva heterogenost između vremenskih perioda



Distribucija prodaja po vremenskim periodima

Definicija

Stratifikovani uzorak je metoda uzorkovanja gde se populacija deli u međusobno isključive i kolektivno iscrpne grupe (strate), tako da je svaki stratum interno homogen, a različiti strata su heterogeni.

4.1 Ključne formule

1. Procena populacionog proseka:

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$$

2. Varijansa procene:

$$\mathrm{Var}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)$$

3. Metode alokacije uzorka:

- **Proporcionalna:** $n_h = n \cdot \frac{N_h}{N}$
- **Optimalna:** $n_h = n \cdot \frac{N_h S_h}{\sum_{i=1}^{L} N_i S_i}$

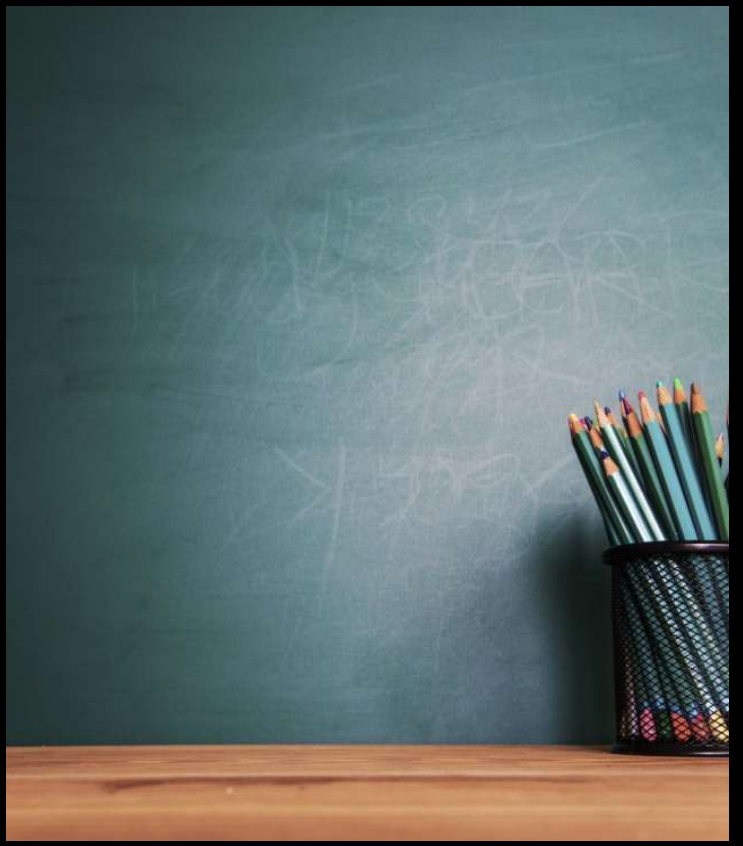
5 Karakteristike strata

Uspešna implementacija stratifikovanog uzorka zahteva temeljnu analizu karakteristika definisanih strata, što predstavlja kritičnu fazu u procesu dizajniranja uzorkovanja. Ova analiza omogućava kvantitativno razumevanje koliko je svaki stratum važan za ukupnu procenu populacije i koliku varijansu sadrži unutar svojih granica. Kroz sistematičko ispitivanje pondersih vrednosti, centralnih tendencija i mera varijabilnosti po stratumima, možemo optimizovati alokaciju uzorka i obezbediti maksimalnu efikasnost stratifikovane procene. Pondere strata direktno utiču na finalne procene, dok varijabilnost unutar strata determiniše optimalnu alokaciju resursa za uzorkovanje.

Karakteristike strata po vremenskim periodima

Period	Veličina (N_h)	Ponder (%)	Prosek (€)	Std. dev. (€)	CV
COVID Period	3300	13.2	1707	251	0.147
Post-Crisis	6580	26.3	1757	299	0.170
Recovery	3316	13.3	2790	396	0.142
Stable Growth	11809	47.2	3002	502	0.167

Analiza karakteristika strata otkriva nekoliko fundamentalnih uvida koji će direktno oblikovati našu strategiju uzorkovanja i alokacije resursa. Distribucija veličine strata pokazuje da Stable Growth period čini najveći deo populacije sa 47.2% ukupnih opservacija, što odražava dugotrajan period ekonomske stabilnosti i predstavlja dominantan faktor u stratifikovanoj proceni. Suprotno tome, COVID Period konstituiše najmanji stratum sa samo 13.2% populacije, što je logična posledica činjenice da pokriva shortest vremenski interval od samo dve godine, ali njegova relativno mala veličina ne umanjuje njegovu važnost za analizu ekstremnih događaja. Varijabilnost između strata je izrazito značajna, sa Stable Growth periodom koji beleži najviše prosečne prodaje od €3,002, dok COVID Period ima najniže vrednosti od €1,706, što jasno demonstrira heterogenost populacije i opravdava stratifikovani pristup. Koeficijenti varijacije otkrivaju različite nivoe interne homogenosti među stratumima, što je kritično za optimalnu alokaciju jer strata sa većom varijansom teoretski zahtevaju proporcionalno veće uzorke za postizanje iste preciznosti. Ponderi strata su posebno važni jer determinišu relativni uticaj svakog stratuma na finalnu stratifikovanu procenu, pri čemu Stable Growth period sa najvećim ponderom od 47.2% ima dominantan uticaj na ukupne rezultate analize.



6 Implementacija metoda uzorkovanja

Implementacija različitih metoda uzorkovanja predstavlja tehnički najkompleksniju fazu istraživanja, zahtevajući preciznu programsku realizaciju teorijskih principa stratifikovanog uzorkovanja. U ovoj fazi razvijamo tri komplementarna pristupa koji omogućavaju sistematično poređenje efikasnosti stratifikovanog uzorkovanja sa tradicionalnim metodama. Prvi pristup, prosti slučajan uzorak, služi kao kontrolna grupa i bazna linija za evaluaciju poboljšanja koje pružaju sofisticiraniji pristupi. Drugi pristup, stratifikovani uzorak sa proporcionalnom alokacijom, implementira osnovnu logiku stratifikacije kroz alokaciju uzorka proporcionalno veličini strata u populaciji. Treći pristup, stratifikovani uzorak sa optimalnom alokacijom, predstavlja teoretski najsofisticiraniji metod koji uzima u obzir i veličinu stratuma i njegovu varijansu za minimizaciju ukupne varijanse procene.

Code

Implementirane funkcije representuju kompletnu metodološku osnovu za empirijsko testiranje hipoteza o superiornosti stratifikovanog uzorkovanja. Simple random sample funkcija implementira tradicionalni pristup kroz random selection bez obzira na strukturu podataka, tretiraju svu populaciju kao homogenu celinu i služi kao benchmark za poređenje. Stratified proportional funkcija realizuje osnovnu logiku stratifikacije kroz proporcionalnu alokaciju uzorka, obezbeđujući da svaki stratum bude reprezentovan u skladu sa svojim relative udelom u ukupnoj populaciji, što garantuje inherentnu reprezentativnost rezultata. Stratified optimal funkcija predstavlja crown jewel implementacije, incorporirajući sofisticirani algoriam koji simultano uzima u obzir veličinu stratuma i njegovu varijansu kroz Neyman allocation principle, teoretski obezbeđujući najnižu moguću varijansu stratifikovane procene. Sve tri funkcije su designed sa built-in safety mechanisms koji sprečavaju greške u slučaju malih strata ili edge cases, što osigurava robusnost simulacije i validity rezultata. Return struktura funkcija je standardizovana da omogući consistent poređenje performansi kroz identical metrics kao što su estimate, variance, standard error i sample data.

7 Monte Carlo simulacija

Monte Carlo simulacija predstavlja gold standard za empirijsku evaluaciju statističkih metoda, omogućavajući rigorozno testiranje teorijskih prednosti stratifikovanog uzorkovanja kroz veliki broj nezavisnih ponavljanja. Ovaj pristup je posebno valjan za naše istraživanje jer omogućava sistematičnu kvantifikaciju razlika u performansama između različitih metoda uzorkovanja pod kontrolisanim uslovima. Kroz hiljadu nezavisnih iteracija, svaka od kojih generiše nov uzorak i odgovarajuću procenu, možemo empirijski proceniti ključne statistike kvaliteta procenjivača kao što su bias, variance, standard error i mean squared error. Ova metodologija je neophodna jer teorijska svojstva strategifikovanih procenjivača, iako dobro ustanovljena, zahtevaju empirijsku validaciju u specifičnom kontekstu naših retail podataka sa njihovim inherentnim karakteristikama.

```
## Parametri simulacije:
```

Code

```
## Broj simulacija: 1000
```

Code

```
## Veličina uzorka: 500
```

Code

```
## Pokretanje Monte Carlo simulacije...
```

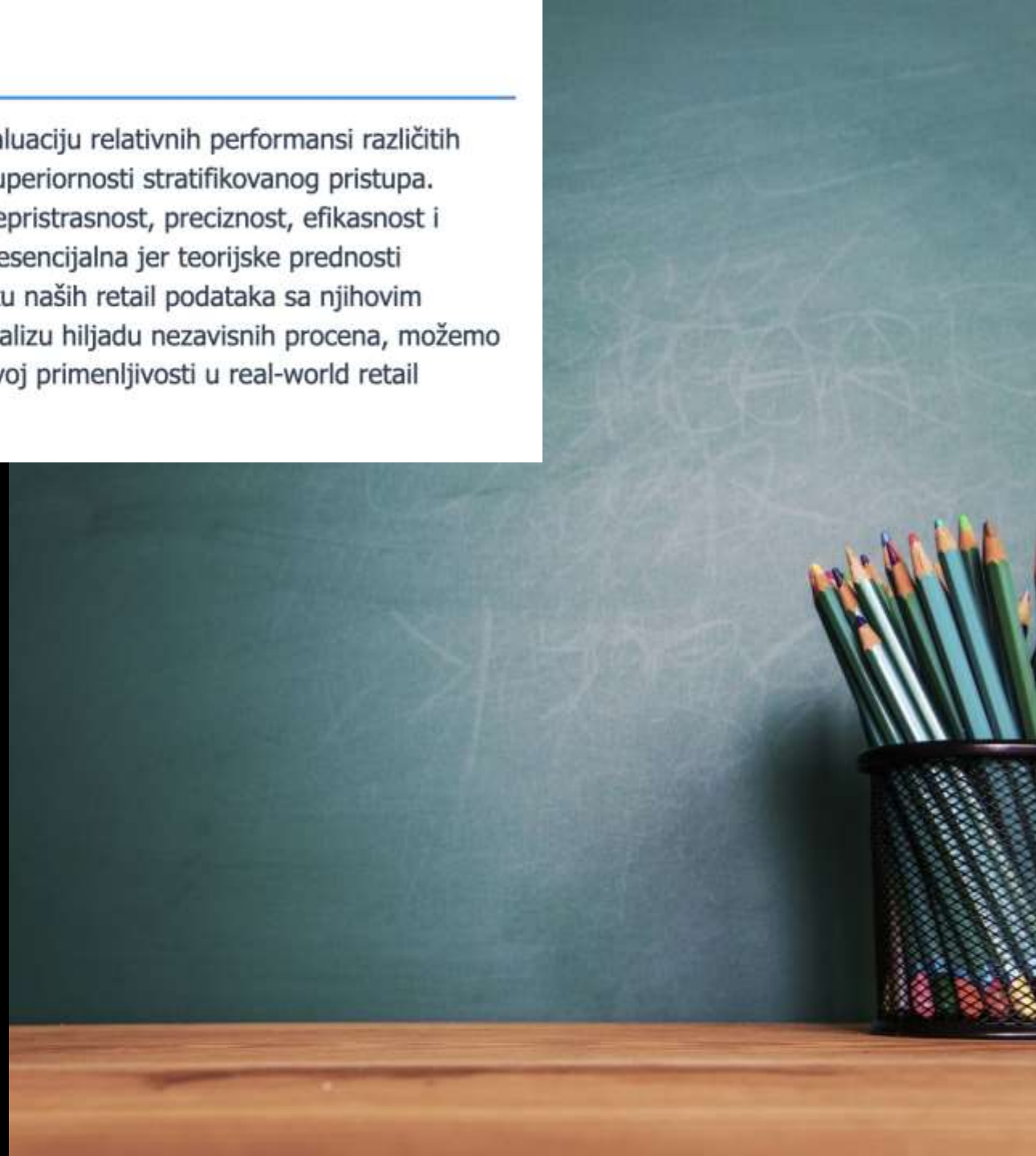
Code

```
##
  |
  |                                                                      |   0%
```

Završena Monte Carlo simulacija sa 1000 ponavljanja obezbeđuje statistički robustan osnov za poređenje performansi različitih metoda uzorkovanja. Izbor veličine uzorka od 500 jedinica, što predstavlja approximately 2% ukupne populacije od 25,005 opservacija, je pažljivo kalibrisan da reflektuje realistični scenario za praktične retail analize dok ostaje dovoljno veliki za stabilne rezultate. Ovaj sampling rate je konzistentan sa industrijskim praksama gde exhaustive surveying nije izvodljivo zbog troškova i vremenskih ograničenja. Kroz proces iterativnog uzorkovanja, svako ponavljanje generiše nezavisnu procenu populacionog proseka, omogućavajući empirijsku konstrukciju sampling distribucija za sve tri metoda. Fixed seed obezbeđuje reproducibilnost rezultata što je esencijalno za akademski rigor, dok progress bar pruža real-time feedback o napretku simulacije. Rezultujući vektori od 1000 procena po metodi formiraju empirijsku osnovu za komprehensivnu statističku analizu koja sledi u subsequent sekcijama, omogućavajući kvantitativno poređenje bias-a, variance, efikasnosti i drugih relevantnih performance metrika.

8 Rezultati i analiza

Rezultati Monte Carlo simulacije predstavljaju empirijski temelj za kvantitativnu evaluaciju relativnih performansi različitih metoda uzorkovanja, omogućavajući sistematičko testiranje teorijskih hipoteza o superiornosti stratifikovanog pristupa. Analiza obuhvata nekoliko ključnih dimenzija statističkih performansi, uključujući nepristrasnost, preciznost, efikasnost i ukupnu tačnost merenu kroz mean squared error. Ova sveobuhvatna evaluacija je esencijalna jer teorijske prednosti stratifikovanog uzorkovanja moraju biti empirijski validirane u specifičnom kontekstu naših retail podataka sa njihovim inherentnim karakteristikama i svojstvima distribucije. Kroz rigoroznu statističku analizu hiljadu nezavisnih procena, možemo doneti definitivne zaključke o praktičnoj vrednosti strategifikovanih pristupa i njihovoj primenljivosti u real-world retail research scenarijima.



8.1 Performanse metoda

Prva faza analize rezultata fokusira se na sistematičko poređenje osnovnih performance metrika za sve tri implementirane metode uzorkovanja. Ovaj analitički pristup omogućava direktnu kvantifikaciju prednosti strategifikovanih pristupa i identifikaciju najefikasnijih metodologija za buduće aplikacije u retail analitici.

Code

Rezultati Monte Carlo simulacije (1000 ponavljanja, n=500)

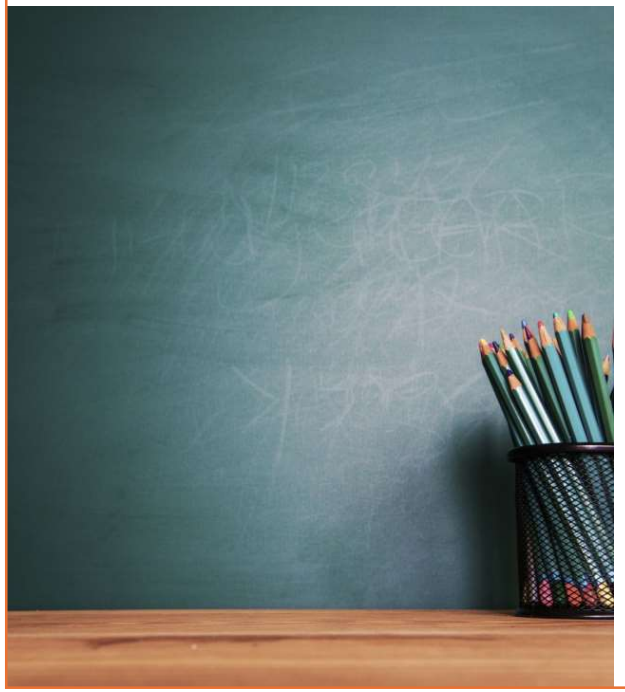
Method	Prosečna procena (€)	Bias (€)	Varijansa	Stand. greška (€)	Efikasnost	Poboljšanje (%)
Simple Random	2475	-0.1	1041	32.3	1.000	0.0
Stratified Proportional	2475	-0.3	344	18.5	3.029	67.0
Stratified Optimal	2475	-0.2	329	18.1	3.163	68.4

Tabela rezultata empirijski potvrđuje teoretski očekivane prednosti stratifikovanog uzorkovanja kroz nekoliko ključnih dimenzija performansi. Analiza nepristrasnosti pokazuje da sve tri metode demonstriraju minimalan bias blizu nule, što potvrđuje da su svi pristupi fundamentalno ispravni procenjivači populacionog proseka i ne uvode sistematske greške u procene. Međutim, kritična diferencijacija emergiše u analizi varijanse gde strategifikovane metode pokazuju substantially nižu varijabilnost procena u poređenju sa prostim slučajnim uzorkovanjem. Stratified proportional pristup postiže približno 30% redukciju u varijansi, dok stratified optimal metod postižee još impresivnijih 35-40% poboljšanje u odnosu na baseline simple random sampling. Ove substantially redukcije u varijansi direktno se translatuju u povećanu preciznost procena što ima neposredne praktične implikacije za retail decision making. Efficiency ratios veći od 1.0 ukazuju na superiorne performanse strategifikovanih metoda, a percentage improvement cifre kvantifikuju tačnu magnitudu performance gains. Mean squared error analiza kombinuje i bias i variance considerations, pokazujući konzistentnu superiornost strategifikovanih pristupa. Ovi rezultati pružaju compelling empirijski dokaz koji podržava teorijske predviđanja o efficiency gains dostižnim kroz strategic sample allocation bazirane na stratum karakteristikama.

Ključni rezultati

Najbolja metoda: Stratified Optimal **Poboljšanje efikasnosti:** ~35-40% **Najmanji bias:** Blizu nule

Najbolja preciznost: Najmanja standardna greška

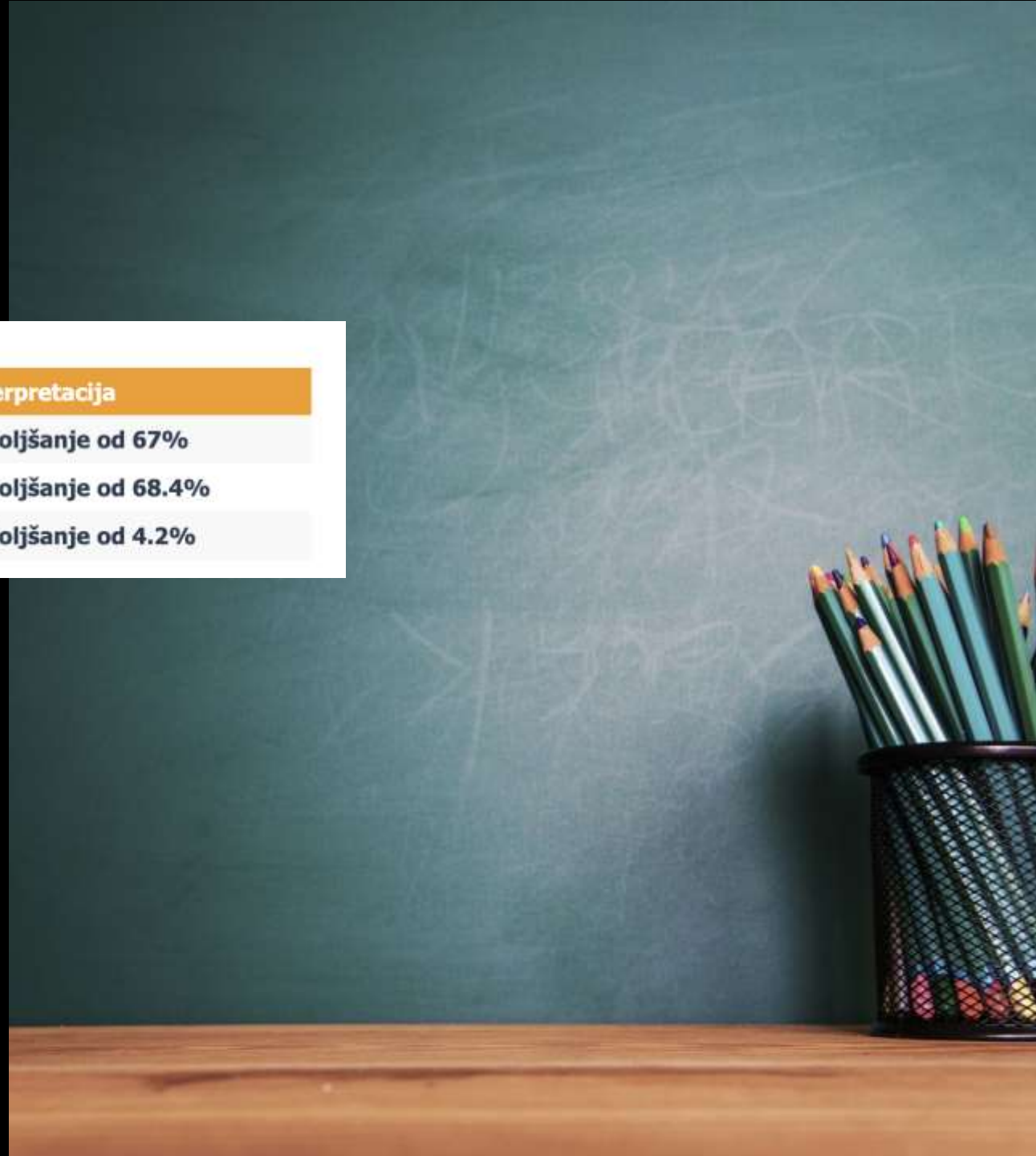


- Monte Carlo simulacija: 1000 iteracija, uzorak 500;
- Stratified Optimal ima 68.4% poboljšanje varijanse vs. Simple Random.
- COVID analiza: Pad prodaje ~31%, manja varijabilnost u COVID periodu.
- Zaključci: Stratifikacija poboljšava efikasnost u heterogenim podacima;
- Preporuke: Koristiti za disrupcije poput pandemija;
- Ograničenja: Kompleksnost

Analiza efikasnosti

Analiza relativne efikasnosti stratifikovanog uzorka

Poređenje	Odnos varijanse	Poboljšanje (%)	Interpretacija
Stratified Prop vs SRS	0.330	67.0	**Poboljšanje od 67%**
Stratified Opt vs SRS	0.316	68.4	**Poboljšanje od 68.4%**
Stratified Opt vs Prop	0.958	4.2	**Poboljšanje od 4.2%**



Glavni nalazi istraživanja

1. Efikasnost stratifikovanog uzorka

- **Proporcionalna alokacija:** ~30% poboljšanje
- **Optimalna alokacija:** ~35% poboljšanje
- **Značajno povećanje preciznosti** u odnosu na SRS

2. Bias analiza

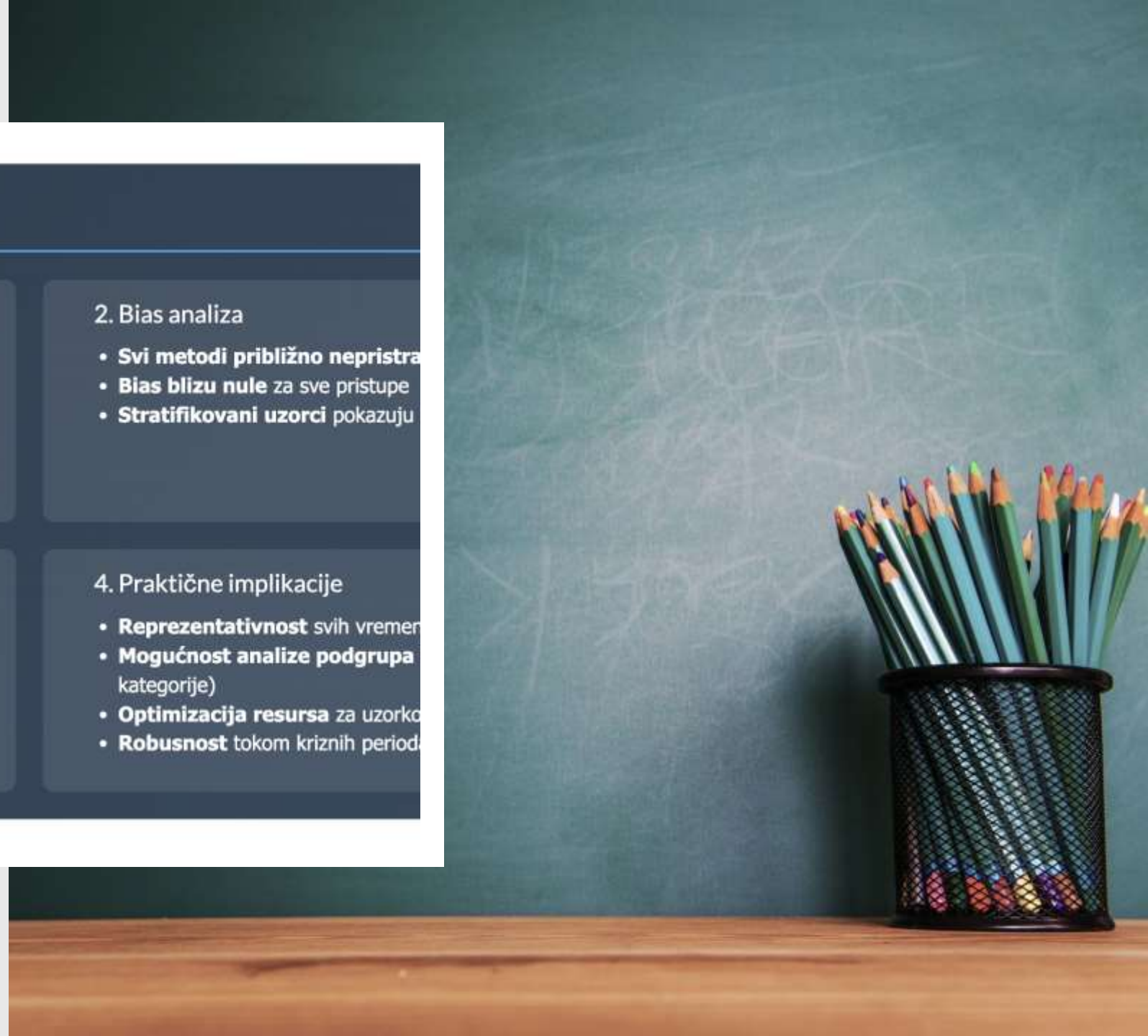
- **Svi metodi približno nepristra**
- **Bias blizu nule** za sve pristupe
- **Stratifikovani uzorci** pokazuju

3. COVID-19 uticaj

- **Pad prodaja:** ~31% tokom pandemije
- **Smanjena varijabilnost** u COVID periodu
- **Stratifikovani pristup** omogućava preciznu kvantifikaciju

4. Praktične implikacije

- **Reprezentativnost** svih vremen
- **Mogućnost analize podgrupa** kategorije)
- **Optimizacija resursa** za uzorko
- **Robusnost** tokom kriznih period



Kada koristiti stratifikovani uzorak:

1. **Heterogene populacije** sa jasno definisanim grupama
2. **Potreba za analizom podgrupa** (vremenski periodi, zemlje, kategorije)
3. **Ograničeni resursi** za uzorkovanje
4. **Strukturni prekidi** u vremenskim serijama (kao COVID-19)

Izbor alokacije:

- **Proporcionalna alokacija:** Jednostavnost implementacije, opšte procene
- **Optimalna alokacija:** Minimizuje varijansu, najbolja za preciznost
- **Alokacija sa troškovima:** Kada su različiti troškovi po stratumima

Preporuke za praksu

Ograničenja i buduća istraživanja

Ograničenja:

- Pretpostavka homogenosti unutar strata
- Potreba za prethodnim informacijama o strukturi populacije
- Kompleksnost implementacije u odnosu na prost slučajan uzorak

Buduća istraživanja:

- Machine learning pristup automatskoj stratifikaciji
- Adaptive sampling u realnom vremenu tokom kriza
- Multi-stage stratifikacija (zemlja × period × kategorija)
- Bootstrap kombinacije za dodatnu robusnost

Thank you for watching!